Unsupervised Learning: Validation

Exploratory vs. Confirmatory Analysis

Confirmatory Analysis:

- Seeks to test an a priori hypothesis.
- Examples:
  - Classical inferential statistics.
  - Prediction in supervised learning.

Exploratory Analysis:

- Seeks to explore and understand data.
- Hypothesis generating.

Which is unsupervised analysis?

Unsupervised Learning Objectives

1. Data exploration.
   - Explore data to find patterns and generate hypotheses.
   - Exploratory analysis - Generated hypotheses will be later tested and confirmed.
2. Data-driven discoveries.
   - Make discoveries by finding important patterns and structures in the data.
   - Challenge: Both exploratory & confirmatory analysis!

Data Exploration

- Understand patterns, trends, and anomalies in your data.
- Prepare your data for the modeling stage and confirmatory analysis.
- Give a visual summary of the data.
- Generate hypotheses to be confirmed later.

We need to use multiple techniques to fully explore and visualize data!

From Data Exploration to Data-Driven Discoveries

Challenge:

- Any potential discoveries found during data exploration (hypothesis generation) cannot be confirmed on the same data set (hypothesis testing).

Why not?

From Data Exploration to Data-Driven Discoveries

Challenge:

- Any potential discoveries found during data exploration (hypothesis generation) cannot be confirmed on the same data set (hypothesis testing).

Why not?

Overfitting!

From Data Exploration to Data-Driven Discoveries

Challenge:

- Any potential discoveries found during data exploration (hypothesis generation) cannot be confirmed on the same data set (hypothesis testing).

Why not?

Overfitting!

What can we do to validate data-driven discoveries?

Validating Data-Driven Discoveries

1. Corroborate via existing literature.
    - Suggestive, but further confirmation needed.

Validating Data-Driven Discoveries

1. Corroborate via existing literature.
    - Suggestive, but further confirmation needed.
2. Show data-driven discovery is stable.
    - Small changes to the data, the algorithm, the method, the parameters, etc. yield the same or similar result.
    - Communicates uncertainty in discovery and avoids overfitting.

Validating Data-Driven Discoveries

1. Corroborate via existing literature.
   - Suggestive, but further confirmation needed.
2. Show data-driven discovery is stable.
   - Small changes to the data, the algorithm, the method, the parameters, etc. yield the same or similar result.
   - Communicates uncertainty in discovery and avoids overfitting.
3. Confirm via a completely separate test set.
   - Strong validation and standard in machine learning.
   - Sometimes challenging for unsupervised learning.

Validating Data-Driven Discoveries

1. Corroborate via existing literature.
    - Suggestive, but further confirmation needed.
2. Show data-driven discovery is stable.
    - Small changes to the data, the algorithm, the method, the parameters, etc. yield the same or similar result.
    - Communicates uncertainty in discovery and avoids overfitting.
3. Confirm via a completely separate test set.
    - Strong validation and standard in machine learning.
    - Sometimes challenging for unsupervised learning.
4. Validate via biological experiments.
    - Gold Standard! True confirmation.
    - Expensive & sometimes not possible.

Confirming Discoveries on a Test Set

Idea:

- Use a "training set" for data exploration to make discoveries.
- Use a separate "test set" for confirming the discovery.

Approach:

- Use a separate study for a test set.
- Randomly split samples into a separate training and test set before any analysis.

Important:

- Keep test set hidden until confirmatory stage!

But, this is often challenging for unsupervised learning ....

Discussion

How would you use a separate test set to confirm the following types of data-driven discoveries'?

- Pattern.
- Clusters.
- Selected features.

Stability Principle

Idea:

- Reproducible discoveries are likely to be true discoveries.

Approach:

- Use repeated data perturbations of training data to mimic having new test data.
  - Subsampling, bootstrapping, random corruptions, random noise, random tuning parameters, random initializations, etc.
- Apply machine learning technique to each perturbed data set.
- If the same discovery appears repeatedly, then it's likely a true discovery.

Stability Principle

Idea:

- Reproducible discoveries are likely to be true discoveries.

Approach:

- Use repeated data perturbations of training data to mimic having new test data.
  - Subsampling, bootstrapping, random corruptions, random noise, random tuning parameters, random initializations, etc.
- Apply machine learning technique to each perturbed data set.
- If the same discovery appears repeatedly, then it's likely a true discovery.

- Useful when a separate test set in unavailable or difficult to confirm via a test set.
- Communicates level of uncertainty associated with the data-driven discovery.

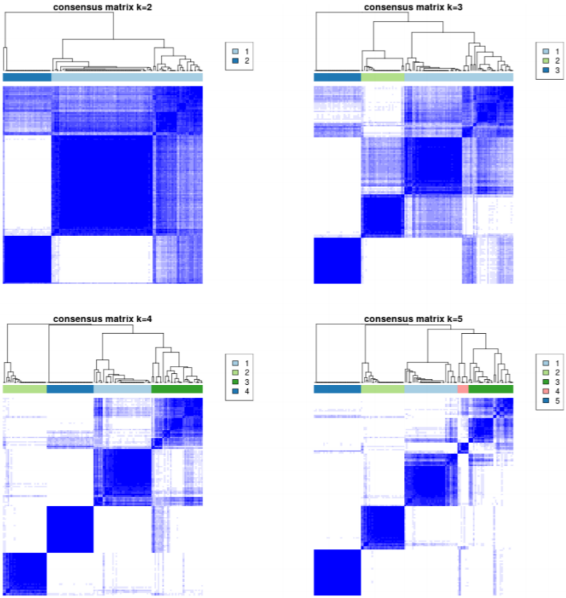
Discussion

How would you use the stability principle to confirm the following types of data-driven discoveries?

- Patterns.
- Clusters.

How would you randomly perturb your data to perform stability analysis?

Stability Example - Consensus Clustering



Unsupervised Learning: Best Practices

Summary

Topics & Techniques covered:

- Dimension Reduction.
  - PCA.
- Pattern Recognition & Data Visualization.
  - PCA, NMF, ICA, MDS, tSNE, UMAP.
- Clustering.
  - K-means, Hierarchical, Biclustering, Convex clustering.
- Feature Filtering via Association Testing.
  - FWER, FDR, Permutation approaches.
- Graphical Models.
  - Graph types, Gaussian graphical models.
- Validation.

Some Good Rules for Unsupervised Analysis

1. Always visualize.
2. Use multiple techniques.
3. Validate discoveries.
4. Communicate uncertainty.
5. Make your analysis reproducible.